Tampere University

Syed Mustufain Abbas Rizvi

Attribution Modelling of Online Advertising

Faculty of Information Technology and Communication Sciences
M.Sc. Thesis
April 2019
Supervisors: Kati Iltanen and Martti Juhola

ABSTRACT

Rizvi, Syed Mustufain Abbas: Attribution Modelling of Online Advertising
M.Sc. Thesis, 62 pages
Tampere University
Faculty of Information Technology and Communication Sciences
April 2019

Attribution modelling is one of the most sought-after research topics in digital marketing. While much research and progress has been made into predictive modelling, attribution modelling requires abundant domain expertise and interpretability for it to be adopted and used by marketers. Many approaches have been laid out including logistic regression and graph-based attribution models such as markov chain which has shown consistent results while retaining high interpretability.

In this thesis, we work on a data set which includes logs of user activity such as user clicks, impressions, and user conversions. The thesis makes use of two different kinds of analysis, user level and sequence level in which different logistic regression models and markov chains are used to assess the performance of attribution on a varied set of metrics and address the class imbalance problem which frequently occurs with user log data.

Keywords: Logistic regression, Markov chain, Google Analytics, Adform, Digital marketing

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

PREFACE

I would like to take the opportunity to thank Dr. Nissanka Wickremashinghe (Data Scientist at Annalect Finland) for guiding me through the process and providing me with the necessary domain knowledge required for this project. His constructive criticism helped me to shape my project and deliver quality work. I'd like to thank my supervisors Martti Juhola (Professor, Tampere University) and Kati Iltanen (Vice Dean, Faculty of Information Technology and Communication Sciences, Tampere University) for their support, guidance, and leniency.

I would like to thank my fiancé for the constant support, always pushing me through and last but definitely not the least, I am immensely grateful to my parents for all the guidance and freedom to allow me to pursue my dreams.

Once again, I'd like to express my gratitude to all the people mentioned above, without them, I would not have been able to execute this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Ad	Advertisement
AUC	Area under Curve
CI	Confidence Interval
CR	Conversion Rate
CTR	Click through rate
DA	Dominance Analysis
ETL	Extract Transform Load
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GA	Google Analytics
IAB	Interactive Advertising Bureau
LR	Logistic Regression
MLE	Maximum Likelihood Estimation
R^2	Correlation Squared
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
ROI	Return on investment
RW	Relative Weight Analysis
SMOTE	Synthetic Minority Over-sampling Technique
T2C	Time to conversion
TP	True Positive
TN	True Negative
TPR	True Positive Rate
URL	Uniform Resource Locator
V-M	Variability Measure

1. INTRODUCTION

In the era of digital technology, social media and with the ever-increasing use of the internet, digital marketing has been on the rise. According to a survey conducted by IAB, Online Advertising Revenue accounted for $88 billion in the United States as shown in Figure 1.1. This gives marketers a profound opportunity. [PwC, 2018]

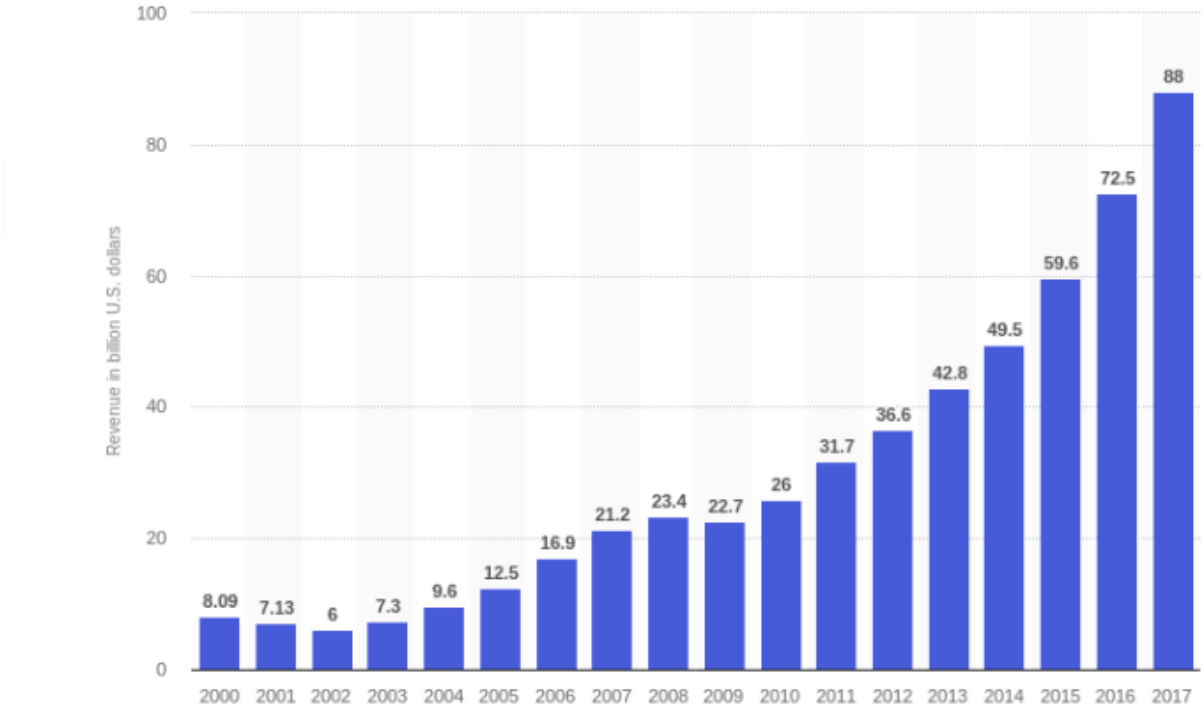

Figure 1.1 Internet advertising revenue report [PwC, 2018]

The most common form of advertising is Display Advertising. Display Ads or the banner Ads are the basic ad unit which is usually embedded within a site, application or game and linked to an advertiser's website. Display Advertising plays a vital role in digital marketing. It can be used to increase sales or redirect users to buy a product. The Internet provides several different channels for online advertising including Email, Social Media, Display and Paid Search as shown in Figure 1.2.

SHARE OF GLOBAL ADVERTISING BY DIGITAL SUB-CATEGORY, 2016-18 (%)

GLOBAL % SHARE OF ADVERTISING SPEND*

	2016a	2017f	2018f
Display**	11.4 (11.4)	12.2 (12.3)	12.7 (13.0)
Online Video**	3.7 (3.7)	4.8 (4.6)	5.7 (5.6)
Social Media**	1.5 (1.4)	1.8 (1.7)	2.2 (2.0)
Paid Search**	13.3 (12.7)	14.6 (13.6)	15.9 (14.6)
Classified**	3.4 (3.3)	3.5 (3.4)	3.6 (3.5)

Figures in brackets show our previous forecasts from June 2017

Figure 1.2 Advertising spend [Manjur, 2018]

Internet makes uses of cookie to track user activity on a website. A cookie is a unique identifier which is assigned when a user browses on a webpage. Advertisers use cookies to monitor the number of ads that have been shown to a visitor. Website uses cookies to gauge the number of unique visitors and what people do on their website. This can enable marketers to track user journey across different web pages and on the internet. This information can be combined with third-party data such as Google Analytics (GA) to analyze the purchases which the user has made. These are called conversions. Conversion is the measure of the number of times that a defined action has been taken on the site and successfully linked to a previous creative impression or click e.g. downloading a brochure or making a purchase. Click is a user action such as clicking an ad with their mouse or touching the screen of a mobile device, which sends them to a click-through URL, indicating interest or engagement. An impression is a single display and view of an ad on a webpage.

In digital marketing, the most common problem is to optimize the marketing spend in such a way as to achieve the highest number of conversions. Machine learning has provided an opportunity to do data-driven marketing and optimize the marketing spend while achieving the highest sales. In traditional marketing, the last click model was used to optimize marketing spend and effect of the different campaign along with a consumer path. This model had serious pitfalls as it assumed that, along a consumer journey, the last touch point was the most significant and got 100% of the credits, thus ignoring the rest of the touch points along a consumer path leading up to a conversion. Machine learning tends to consider all the touch points along a consumer journey and assign credits to them by learning from data, thus producing more reliable results. It also addresses the class imbalance problem as there are very few conversions. The aim of this research is to use logistic regression and markov chains to attribute conversions to different marketing channels. The methods used should be easy to interpret while providing state of the art results.

The thesis is organized as follows. Chapter 2 describes attribution modelling and the methods that are being used for assigning credits to conversions. Chapter 3 talks about different areas of machine learning and its methodologies. Chapter 4 discuss the research that has been carried out in recent years in the area of attribution modelling. Chapter 5 describes data and gives a detailed exploratory and descriptive data analysis on said data. Chapter 6 explains the methods that were used while conducting the research. Chapter 7 gives detailed results of the experiments. Chapter 8 is the conclusion.

2. ATTRIBUTION MODELLING

Attribution modelling is the set of rules that determines how the credit for sale and conversions is allocated to the touch points in conversion paths. Conversions paths are those paths along which conversion has taken place. Touch point can be defined as a user interaction with a business through a website or any other application. In marketing, it defines the ways information is displayed to the prospective user. Attribution modelling maps the user journey from conversion to user considering every touch point along the way. It helps marketers to see the impact of the different touch points in user conversion and thus can be used to optimize their marketing spend, obtaining the highest rate on investment (ROI) and conversions e.g. 80% of the conversions come from Display Ads while the rest come from Social Media. This would help marketers to invest more in Display Ads as compared to Social Media marketing which would help in increasing their ROI.

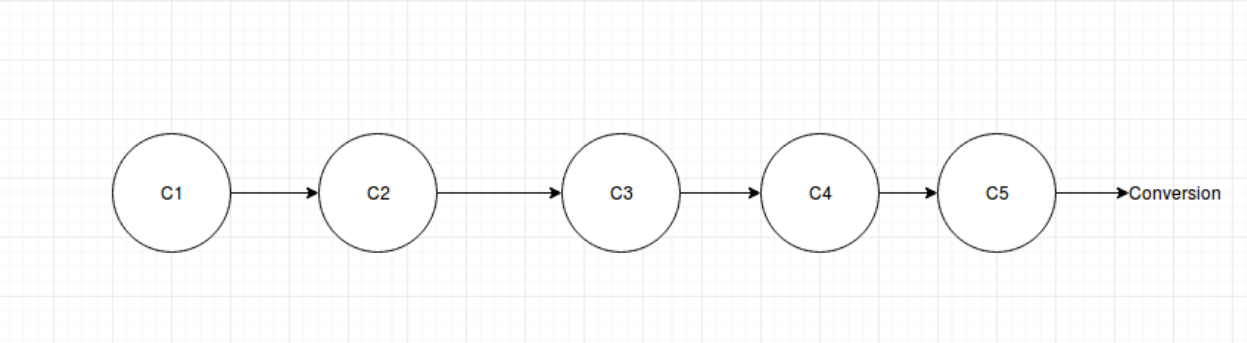

Figure 1.3 Converting journey with 5 touch points C1, C2, C3, C4 and C5

2.1 First-Touch Attribution Model

First touch attribution model assigns all the credits to the first touch point in the journey. For example, in Figure 1.3, 100% of the credits will be assigned to C1. However, it has a serious pitfall and only conveys a part of the story as it ignores the rest of the touch points. An advantage of this model is that it is easy to implement and easily interpretable for marketers. [Con, 2016]

2.2 Last-Touch Attribution Model

Last touch attribution model assigns all the credits to the last touch point in the journey. For example, in Figure 1.3, 100% of the credits will be assigned to C5. It is the most popular model that is used by marketers. As the first touch, it is simple to implement but only considers the last touch point leading up to conversion and thus only gives an idea of what happens at the end of the journey. [Con, 2016]

2.3 Linear Attribution Model

The linear model assigns credits equally between touch points along the journey. For example, In Figure 1.3, credits would be distributed equally among these five touch points. Each touch point would receive 20% of the credits. It is better than the last touch and the first touch as it explains the whole journey and not just one touch point. However, it is not true in all cases that each touch point contributes equally to a conversion. [Con, 2016]

2.4 Time Decay Attribution Model

Time decay attribution model assumes that touch points that are closer to conversion along a journey contribute more and hence receive more of the credits. For example, in Figure 1.3, C1 would receive 15% of the credits. C2 17%, C3 19%, C4 23% and C5 26%. It considers multi-touch points as opposed to single touch points and it makes sense that touch points closer to conversion would have more impact on it. However, in some cases, initial touch points might have a greater impact which this model tends to ignore. [Con, 2016]

2.5 Position-Based Attribution Model

Position based attribution model, also called U-shaped attribution model, assigns 40% of the credits to the first touch point, 40% credits to the last touch point and 20% of the credits are evenly distributed among the rest of the touch points. It addresses several flaws from the previous model by placing emphasis on the middle part of the journey and still considers the contribution of the first and the last touch point. For example, in Figure 1.3, 40% of credits would be assigned to C1 and C5. C2, C3, and C4 would get 6.7% of credits. However, it is not necessary that the touch points in the middle part of the journey contribute evenly to conversions. [Con, 2016]

2.6 Algorithmic Attribution

Algorithmic attribution [Anderl *et al.* 2013] is by far the most advanced technology used in attribution modelling. It uses machine learning techniques to allocate credits to different touch points along a journey. It uses historical customer data to run machine learning algorithms and assigns weights to different touch points by learning from data.

3. MACHINE LEARNING

Machine learning is a field of computer science that uses statistical techniques to allow computers to improve their learning over time by feeding more and more data. Arthur Samuel came up with the term *machine learning* in 1959 [Samuel, 1959] and is regarded as the pioneer of this field. It makes use of historical data to learn from it and provide inference on unseen data which the model has not seen before. Machine learning can be widely divided into two broad categories, supervised and unsupervised learning.

Supervised learning: It makes use of data with labels. The main objective is to learn a function that maps input to output based on input-output pairs [Russel and Norvig, 2009]. In supervised learning, each example is a pair of input object which is data and output object which is a label.

Unsupervised learning: It makes use of data without labels. The main objective is to infer a function that describes the structure of unlabeled data.

Classification: It is a supervised learning technique where the classes are discrete. The data consists of two or more classes. In a classification task, machine learning algorithms learn the underlying decision boundary that separates the two classes into two regions and then assigns the unseen data to a class as shown in Figure 1.4.

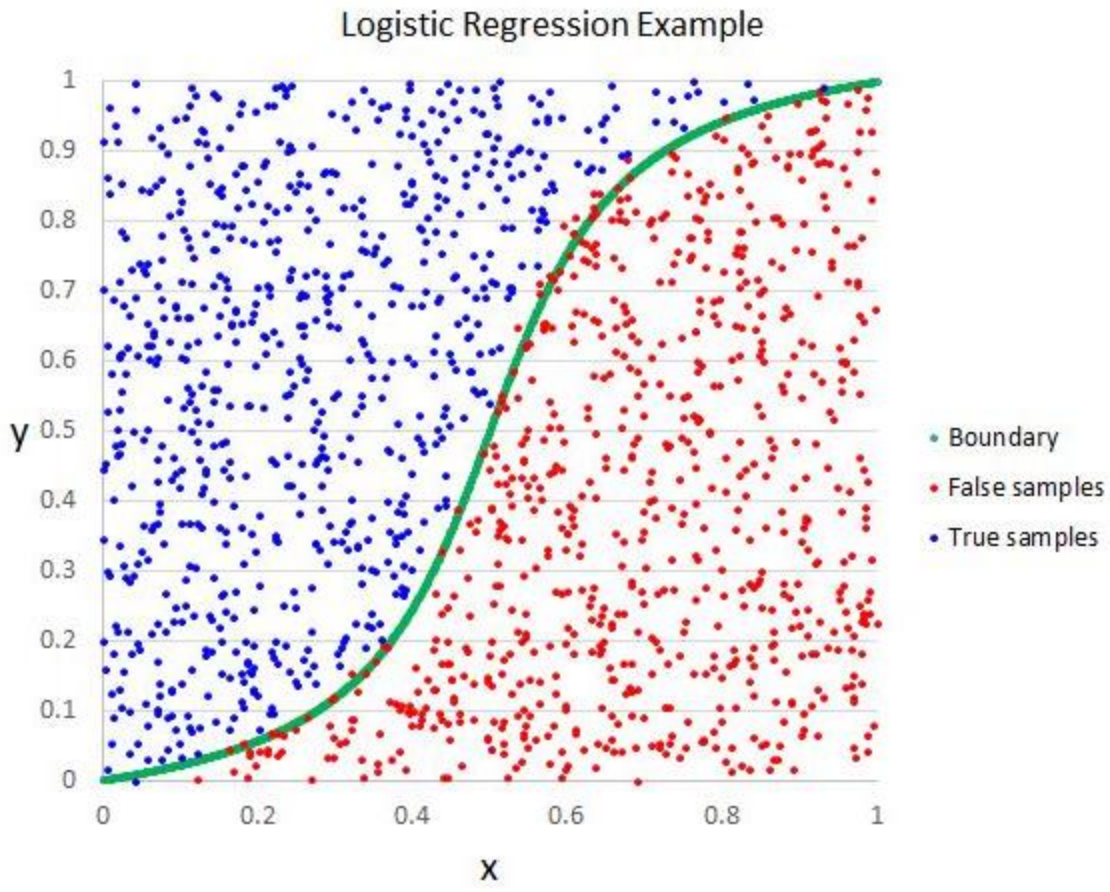

Figure 1.4 Classification. [ACMer, 2016]

Clustering: It is an unsupervised learning technique. Clustering divides the data into *k* number of clusters. It puts similar data points into the same cluster using different clustering algorithms.

Regression: The data has continuous labels. It is also a type of supervised learning

Machine learning nowadays is making an impact in every walk of life. With the ever-increasing use of the internet and rise of digital technology, machine learning has made a significant impact in digital marketing and optimizing marketing budget based on algorithmic attribution. This enables marketers to do data-driven marketing in an efficient way as compared to traditional attribution models. The results have consistently improved over the years and have provided reliable insights to marketers.

3.1 Logistic Regression

Logistic regression is a supervised learning technique [Cox, 1958]. It is a probabilistic model where it outputs probabilities based on a sigmoid function as shown in Figure 1.5. It defines the relationship between dependent variable also called label which is usually binary and one or more independent variables.

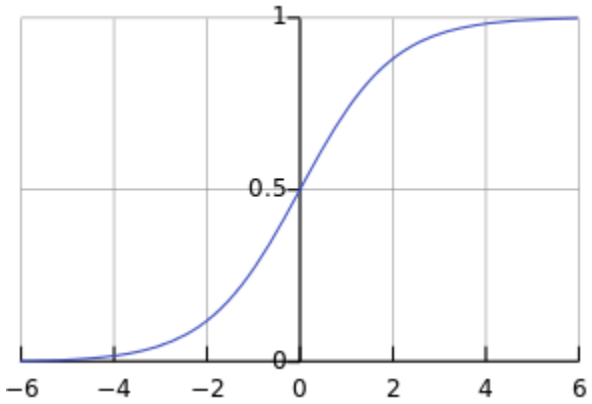

Figure 1.5 Sigmoid function

$$S(x) = \frac{1}{1+e^{-x}}$$

Logistic regression is the natural logarithm of the odds ratio. The odds ratio is defined as the ratio of one odd divided by another. The odds ratio represents the odds that an outcome will occur given a particular event, compared to the odds of the outcome occurring in the absence of that event. Logistic regression can be defined according to the following equation:

$$p(label) = \beta_o + \beta_1 * (Independent\ Variable\ 1) + \beta_2 * (Independent\ Variable\ 2)$$

In terms of odds ratio it can be written as:

$$Log(P/(1-P)) = \beta_o + \beta_1 * (Independent\ Variable\ 1) + \beta_2 * (Independent\ Variable\ 2)$$

If the β_1 value is 1.6, it means that 1-unit change in $Independent\ Variable\ 1$ while others independent variables are at the same level, produces 1.6-unit change in the natural log. In logistic regression β_i are estimated using maximum likelihood estimation (MLE). MLE attempts to find the values of β_i that maximize the likelihood function, given the observations. The resulting estimate is called a **maximum likelihood estimate.** Logistic regression is a very popular machine learning algorithm which is used in many

applications such as fraud detection, image segmentation etc. The ease of interpretation with logistic regression has enabled it to use in a variety of applications. [Peng *et al.* 2002]

3.2 Markov Chain

A markov chain describes a sequence of possible events in which the probability of each event depends only on the previous state. Markov chain was named after Russian mathematician Andrey Markov. A process is said to inhibit markov property when then conditional distribution of future events depends only on the present event and not on the past event. Markov chain is depicted through a directed graph as shown in Figure 1.6. The vertex of the graph is called states. The edges show the probability of moving from one state to another state. A discrete time markov chain has a finite set of states. The markov chain graph represents the state transition probability of *m x n* matrix where *m* is the current state and *n* is the next state. [Powell and Lehe, 2014]

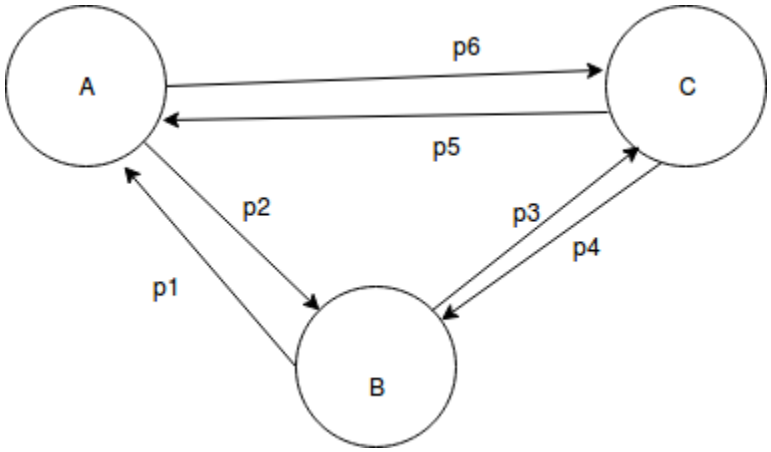

Figure 1.6 A Markov chain with 3 states

Markov chain is used in a variety of application such as weather forecasting [Weiss,1964]. Google PageRank algorithm Page *et al.* [1998] uses markov chain to determine the order of search results. An automated subreddit [Gehl and Bakardjieva, 2016] populated by bots making submissions and comments primarily uses markov chains based on the actual posts in other subreddits. They have also been applied in generating text [Gehl and Bakardjieva, 2016] and modelling game of chance Daykin *et al.* [1967] such as Snakes and ladder, Monopoly etc.

4. LITERATURE REVIEW

There is a considerable amount of research that has been done on attribution modelling and the domain of marketing. Data-driven marketing has been an area of interest for marketers and use of machine learning has proved to give reliable results which are being used by marketers to drive their marketing strategies. The use of probabilistic models and markov chains have provided considerable insights into the area of attribution modelling and given significant results.

One of the biggest challenges with attribution modelling is to retain the interpretability and stability of the model which can be used by marketers to optimize their marketing strategies. One solution proposed by [Shao and Li, 2011] uses bagged logistic regression for attribution modelling. The motivation behind the solution was to come up with a bivariate metric to assess the performance of the attribution model. The research uses two comparable models bagged logistic regression and a probabilistic model. Bivariate metric uses the average misclassification rate and average variability of the estimate to assess the performance of the model over 1000 iterations. Bagged logistic regression uses bagging which combines several weak logistic regression models and the results are averaged over 1000 iterations. Bagging reduces variance in the data and gives stable estimates. To avoid having too few positive samples as compared to negative samples, experiments are conducted with ratio 1:1 and 1:4. A probabilistic model is computed in the following way:

$$P(y|x_i) = \frac{N_{positive}(x_i)}{N_{positive}(x_i) + N_{negative}(x_i)}$$

and pair-wise conditional probabilities:

$$P(y|x_i, x_j) = \frac{N_{positive}(x_i, x_j)}{N_{positive}(x_i, x_j) + N_{negative}(x_i, x_j)},$$

y is a binary outcome having values 1 or 0 denoting user conversion or not, and x_i, *i=1….. p* denotes *p* different advertising channels.

$N_{positive}(x_i)$ and $N_{negative}(x_i)$ denote the number of positive and negative users exposed to channel *i*. $N_{positive}(x_i, x_j), and\ N_{negative}(x_i, x_j)$ denote the number of positive and negative users exposed to channel *i* and *j.* The contribution of channel *i* for each user that has converted is computed in the following way:

$$C(x_i) = p(y|x_i) + \frac{1}{2N_{j\neq i}} \sum_{j\neq i} \{p(y|x_i, x_j) - p(y|x_i) - p(y|x_j)\},$$

$N_{j\neq i}$ denotes a total number of *j's* not equal to *i*. For a particular user, it would be a total number of channels minus one (channel *i* itself). The results from the research are summarized in the Table 4.1.

Model	Variability Measure	Misclassification rate
Logistic Regression	2.115	0.091
Bagged Logistic Regression	0.672	0.093
Probabilistic Model	0.026	0.115

Table 4.1 Results of the Experiment. **[Shao and Li, 2011]**

Probabilistic model achieves the lowest variability in the estimates due to its deterministic natures but suffers from high misclassification rate. On the other hand, bagged logistic regression achieves a lower variability measure as compared to normal logistic regression providing more stable estimates.

Regression methods have also been extensively studied in the context of attribution modelling. Zhao *et al.* [2018] have proposed relative importance methods such as dominance analysis and relative weight analysis which aims at allocating coefficient of determination of regression models as attribution values. Relative importance methods choose a regression model that best fits the underlying relationship between revenue and advertising effort and decomposes the resulting R-squared (R^2). In general, R^2 is a statistical measure of how close the data are to the fitted regression line. Dominance analysis and relative weight analysis have been used for R^2 decomposition. Dominance analysis ensures all interactions are fully considered for calculating attribution values by comparing coefficients of all nested sub models which are composed of subsets of independent variables with that of full models. Dominance analysis is computationally expensive as there are $2^p - 1$ submodels to be estimated when computing attribution values. Relative weight analysis, on the other hand, creates a new set of orthogonal variables from original ones which are uncorrelated. These transformed variables can then be used as relative importance values. The first experiment was conducted with 100 data points, generated from a standard linear model which was replicated 30 times. The results showed that dominance analysis (DA) and relative weight analysis (RW) seem to produce quite similar results. They outperformed legacy methods such as regression

coefficients and squared correlations which are also used to assess the importance of the variable. DA and RW were further extended to additive models which incorporate non-parametric model components. The process was repeated again with 5-fold cross-validation and root mean square error (RMSE) was calculated. The results showed that additive models outperformed linear models in terms of predictive accuracy and higher coefficient of determination.

Markovian graph-based approaches have also been used with respect to attribution modelling. A comprehensive graph-based attribution framework has been laid out in Anderl *et al.* [2013]. Six different criteria were used to assess the attribution model in practice, which was Objectivity, Predictive Accuracy, Robustness, Interpretability, Versatility and Algorithmic efficiency. The experiment was conducted on four clicks stream datasets provided by advertisers. There were four variants of the markov model which were treated as simple, forward, backward and bathtub model. In the forward model, states were defined by the channel and position in the customer journey, counted from the beginning. In the backward model the channel and the position in customer journey were counted from the last observation. However, the bathtub model distinguishes these two positions. In total, 16 model variations were evaluated ranging from first-order markov chain to fourth order. For attribution modelling, ad factor removal effect was used which is the change in the probability of reaching the conversion state from the start state when that particular state is removed. Since the data is highly imbalanced, ROC AUC score was used for predictive accuracy. The experiment was conducted using 10- fold cross-validation and predictive accuracy was calculated for both within and out of the sample. The results showed that the second-order Markov chain outperformed the first order while the largest increase in predictive performance was observed from second to third order. The fourth order didn't seem to have a significant increase in predictive performance however it was marginal in most cases. To assess the robustness of the model, ad factor removal effect was evaluated on 10-fold cross-validations. The average standard deviation was then computed for each model state. The results showed that simple and backward first-order markov chain model outperformed the forward model in terms of robustness and with the lowest variation. Robustness tends to decrease as the order of Markov chain increases.

Survival Analysis belongs to the branch of statistics for analyzing the expected duration of time until one or more events happen. Zhang *et al.* [2014] propose one such approach. In terms of survival theory, waiting time is treated as the time it takes for the user to convert. Death is regarded as user conversion. Similar concepts are borrowed to be applied to attribution modelling. They proposed an additive hazard model, which not only consider the contribution of each advertising channel but also the variations of their time decaying speed. Each channel is treated as a set of two parameters, the contribution of

channel and time decaying speed. The time-dependent contribution of each channel is modeled by a hazard function with a set of varied exponential kernel functions which incorporates the contribution of each channel on user conversion. The model is then fitted by maximizing the log-likelihood function in an iterative manner. The experiment was first conducted on synthetic data and relative error was used to measure the accuracy of the parameters inference.

$$\text{Relative error} = \frac{|\beta - \beta^*|}{\beta},$$

where β is a true parameter and β^* is inferred parameter.

The relative error all parameters was 1.4 % which showed that the model performed well on parameter inference. The experiment was further conducted on a real-world data set. Data were split into testing and training dataset on a ratio of 50:50. The data was an imbalance in nature with very few purchases so three metrics were used to assess the predictive power of the mode: precision, recall and F-1 score. The results showed that the model achieved an F1-score of 0.035 and outperformed rest of the models in terms of precision, recall, and F1-score

Comparative analysis of the methods used in online attribution modelling Jayawardane *et al.* [2015] helps to formulate the problem and the research that is been carried out in this domain. The attribution methodologies have evolved over time. They are mainly classified as simplistic or fractional. Simplistic such as *the Last Touch, First Touch* etc. assign complete conversion credit to a single touch point. Fractional attribution models are rules-based models. Instead of assigning conversion credits to a single touch point, they assign to all touch point in consumer journey. These rules are not derived from data rather they are based on intuition or domain knowledge e.g. *Linear, Time decay, Position based* etc. Fractional – Algorithmic uses descriptive modelling approaches to define the underlying relationship between touch points and user conversion. The rules to allocate conversion credits to each touch point in consumer journey are derived from data rather than on heuristics. Logistic regression was considered to be the first attribution model due to its nature of interpretability. Conditional probabilistic models offer a more intuitive understanding but suffer from model accuracy. The causal analysis examines the effect of advertising creative on customer conversions. With regard to multi-touch attribution modelling, the causal framework defines parameters that capture the cumulative marginal uplift created by each touch point. However, in order to produce unbiased causal estimates, date in consideration needs to have some underlying assumptions such as no unmeasured confound and unbiased advertising treatment which is very hard to meet making causal parameters as impractical. In Game theoretical approaches such as shapely value, proposes an approximate method where interpretation is recast as a measure of variable importance. Given the non-variability of all other factors, variable

importance is defined effect on conversion upon changing an exposure or non-exposure to a channel of interest. In the context of Survival analysis "death" denotes a customer conversion within the time period and advertising interventions are defined as hazards. Markov chain, on the other hand, offers more understanding due to state transition probabilities. The consumer journey can be interpreted as a markov process of order n. It takes into account both the conversions and non-conversion. Hidden markov model explicitly incorporates the effect of a preceding exposure (user interaction with a channel), high order markov chains yield higher model accuracy. Bayesian inference on the other hand attempt to capture the dynamic effects between advertising channels, allowing touch point data to be indeterministic stochastic events that trigger subsequent visits by other channels.

5. DATASET AND FRAMEWORKS

5.1 Dataset

The dataset comes from two data sources GA and Adform.
Adform: This is on the advertiser side. Adform data capture each transaction based on its type. Each transaction registered by Adform is captured and placed into Tab Separated Values (TSV) files based on its type. The following are types of files:

- Clicks
- Impressions
- Campaign

Each file contains information that is specific for the transaction. New files are prepared every hour. Size of the files and data volume per day vary and fully depend on user activity and traffic load related to campaigns the agency/advertiser is running. An advertising campaign is a series of advertisement messages that share a single idea and theme which make up an integrated marketing communication. Clicks and impressions are collectively called events.

GA: This is Google Analytics which is on the website side. It shows user conversion when a user goes onto the website to buy a product. It contains column 'transaction' which has two values 0 and 1. 1 means conversion and 0 means non-conversion. It is a binary variable which will be used as a label for the machine learning process.

5.1.1 Data preparation

Raw data is extracted from Adform API and GA API. Adform and GA data are collected over a week. Raw data is then transformed through extract, transform and load (ETL) operations which can then be consumed by the modelling process. Two sets of data are prepared for two types of analysis: **user level analysis** and **sequence level analysis.**

5.1.1.1 User Level data

For the user level analysis Adform impression, Adform click, and Adform campaign data is merged based on cookie id. GA data is then merged with the above data to get user conversion. Click and impression interaction is counted for each distinct user with each and every campaign. From now on we will use campaign and channel interchangeably.

Clicks and impression campaigns are treated separately as it models the problem more efficiently. It is then merged with conversion data to match users onto conversion. The data up to this stage looks in the following format:

Cookie id	A_1	A_2	B_1	B_2	conversion
User A	0	1	3	0	0
User B	0	0	0	0	0
User C	1	10	0	0	1
User D	8	0	2	9	0

Figure 5.1 Input data for User Level Analysis

In Figure 5.1, A and B represents different campaigns. _1 represents clicks campaign and _2 represents impression campaign. Each row represents the journey of a distinct user that is converted or not converted.

After data is prepared it then goes into the preprocessing phase. Sometimes it happens that a user has disabled the cookies in the browser, so cookie id is missing in that case. Such events are removed from the data as there is no way to track those users across different channels. Duplicate cookie id is also removed as each row represents one unique journey of a customer. Campaigns which have all zeros in it, or which have only one distinct value are removed as there is no variance in the data and does not hold any useful information regarding the conversion. Events which happen after the user has converted are removed for that particular user. One drawback of this approach is that some users can have multiple conversions and thus it tends to ignore the multiple conversion per user and only considers whether the user has converted or not. It is further addressed in sequence level analysis

5.1.1.2 Sequence Level data

For sequence level analysis Adform clicks, Adform impression, Adform campaign, and GA data are merged together based on cookie id. Channel pathways are then constructed for each user. In this analysis events where cookie id is missing are removed as there is no way to construct a consumer journey for such users.

For channel path construction we divide the user into three broad categories.

1. Users who have one conversion
2. Users who do not have conversion
3. Users who have multiple conversions

The channel path is constructed for the time window specified. Since data collection is done for one week so channel path for users is constructed for a week. Paths should be constructed in such a way that there should be one conversion per sequence. For users who have one conversion, it is fairly simple to group those users and create the path in ascending order of time such that earliest channel interaction with the user should be first in the sequence and then leading up to a conversion. For users who do not have a conversion path would be constructed in a similar fashion but leading up to non-conversion. For users who have multiple conversion and post-conversion events, they are handled separately as in Figure 5.2.

Figure 5.2 User with multiple conversions

Since each sequence must contain one conversion for each user, a multiple conversion journey is divided into sub journeys for that user as illustrated below in Figure 5.3.

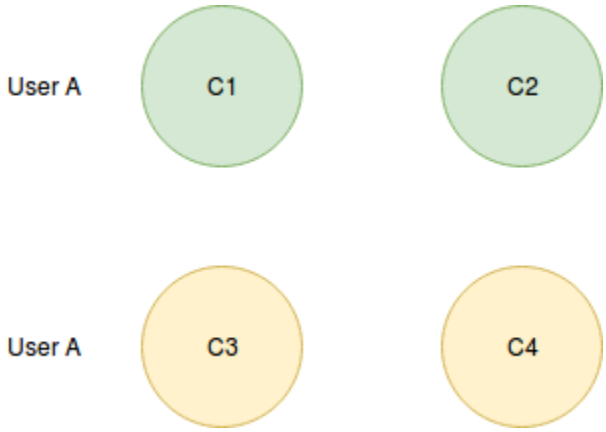

Figure 5.3 User with sub journeys

In this way, sequences are then generated for all the users in that time window leading up to a conversion or not. It also takes into effect the multiple conversions for users by operating on the sequence level which was ignored in user level data. The data up to this stage would be in the following format.

Channel path	Conversion
A_2>A_1>A_1>B_1>B_1	1
B_2>B_2	0
A_2>B_2>	0
A_2>B_1>B_1	0

Figure 5.4 Input data for Sequence Level Analysis

In Figure 5.4, A and B are different campaigns. _1 represents clicks campaign and _2 represents impression campaign. Each row represents the consumer path of that is converted or not converted. The campaigns are in chronological order of user interaction such that the earliest campaign user interacts with that comes first. These are also called touch points.

5.2 Exploratory Data Analysis

Exploratory data analysis included basic visualizations to understand the concept of consumer journey better.

Figure 5.5 Conversion density

In Figure 5.5, the distribution of conversions is analyzed over weekdays. The deeper purple colors represent more user conversions on that day and time. The spaces which are white have zero user conversion on that day and time.

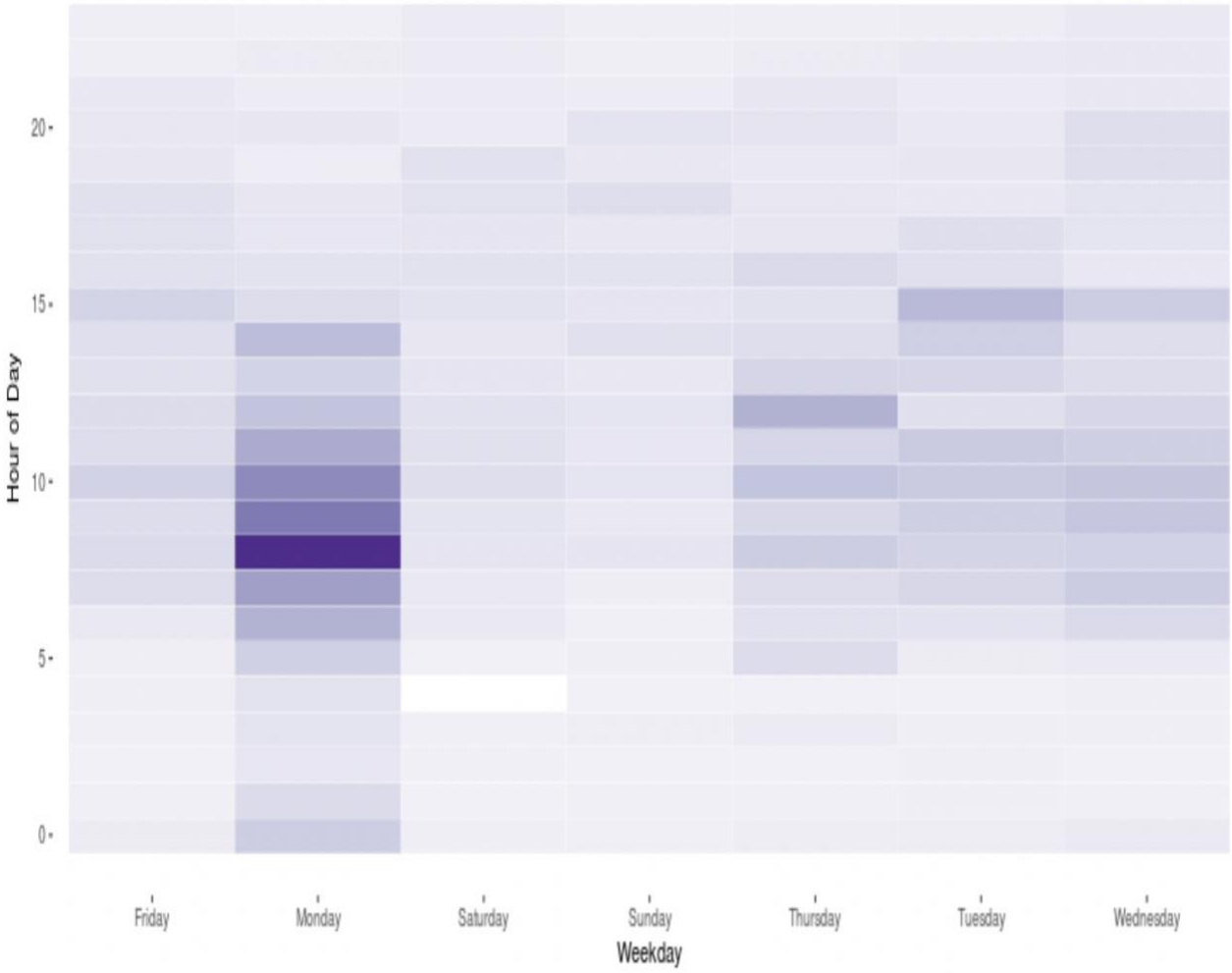

Figure 5.6 Event density

In Figure 5.6, events which are clicks and impressions combined are analyzed. The deeper purple colors represent more user activity on that day and time. As the color starts to fade, it represents lesser user engagement.

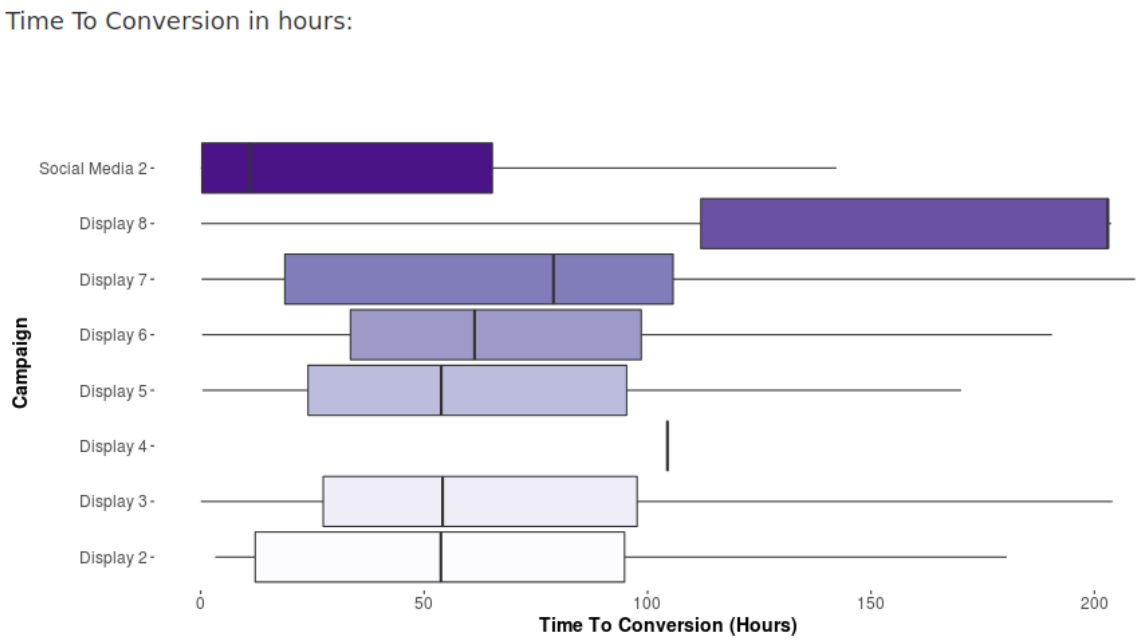

Figure 5.7 Time to the conversion of each channel

In Figure 5.7, boxplots for each channel are constructed which show how much time it takes when a user interacts with these channels and it gets converted. The boxplot shows minimum time to conversion (T2C), lower 25 % of the T2C which is called the lower quartile, median T2C and upper 25% of the T2C which is called the upper quartile. It can be used to effectively see the performance of each of the said channels.

5.3 Descriptive Data Analysis

In this analysis, data is summarized to compute basic statistics. A high number of impressions are present in data (4650873) as compared to clicks (15091). Click through rate (CTR) is very low in this case (0.00324). CTR is the click-through rate which is the ratio of click to impressions as how many times users have clicked on an Ad when it was shown. Higher CTR shows us that there have been more clicks on an Ad, and it was relevant to the users.

5.3.1 Funnel Analysis

Funnel analysis involves using a series of events that lead towards a defined goal, like from user engagement in a mobile application to a sale in an eCommerce platform or advertisement to purchase in online advertising [Jansen and Schuster, 2011]. The origin of the term funnel analysis comes from the nature of a funnel where individuals will enter the funnel, yet only a small number of them will perform the intended goals.
During a customer journey, a user may interact with an organization several times using several different **channels**. Each of these interaction instances represents a **touch point** between the customer and the organization. In this context, funnel analysis reveals the performance of each channel by analyzing the various touch points along a consumer journey.

Funnel creation

- For each user, sort events in ascending order by time.
- Number the touch points sequentially in ascending order of time for each user.
- Add total touch points for each user which would be the maximum value of touch point in the user journey.
- Touch point number for a particular user represents the position of that touch point in a user journey.
- Funnel position is classified into 3 categories, first event, mid pathway, and last event.
- If the position of the touch point in a user journey is equal to 1 then it is the first event.
- If the position of the touch point in a user journey is equal to total touch point in a user journey, then it is classified as the last event.
- If the position of the touch point in a user journey does not fall into the first event or last event, it is classified as a mid-pathway.
- Count the funnel position for each touch point.
- Calculate total events by adding a first event, mid pathway and last event for each touch point.
- Calculate average first event, mid pathway, and last event.
- Calculate the proportion of the first event, mid pathway and last event for each touch point in all user journeys by dividing each of the funnel position with an average of the events.

	Campaign	First event	> Mid pathway	> Last event	Total Events
1	Display 1	150	150	0	2.00
2	Display 2	6	285	9	3,077.00
3	Social Media 1	75	150	75	4.00
4	Social Media 2	25	240	35	5,671.00
5	Display 3	6	287	7	43,589.00
6	Display 4	6	287	7	1,001.00
7	Display 5	11	270	19	13,951.00
8	Display 6	6	287	7	17,141.00
9	Display 7	12	278	10	12,712.00
10	Display 8	9	283	8	830.00

Previous 1 Next

Figure 5.8 Funnel Analysis

In Figure 5.8, it shows us that Display 1 has 150 proportion being the first event in all user journeys which means users engaged with this touch point the highest number of times as being the first touch point but have 0 proportion being the last touch point which is the end of the funnel. This indicates the performance of Display 1 as users did not find this effective enough that they would go to the website e.g. to purchase a product. Funnel analysis is valuable in assessing the performance of each channel in greater detail as it tells user engagement with a particular channel at every step of the user journey and does not take a user journey as a whole.

5.3.2 Pathways Analysis

Pathways analysis computes statistics by creating channel pathways for each user. It gives a detailed analysis of the most effective paths over a particular period of time.

Channel path	Conversions	Path length	Uniques	CR
Display 3	66	28.3	853	7.74%
Social Media 2	29	9.2	684	4.24%
Display 5	25	16.6	314	7.96 %
Display 7	24	7.3	670	3.58 %
Display 6	24	12	371	6.74 %
Display 3 > Display 5	6	41.2	66	9.09 %
Display 7 > Display 3	6	10.2	55	10.91 %
Display 7 > Display 5	6	96.6	49	12.24 %
Display 2	5	30.4	61	8.20%
Display 3 > Display 6	5	58.9	37	13.51%

Table 5.1 Channel pathways

In Table 5.1, top 10 channel pathways are shown with respect to conversions, unique users and conversion rate. Unique users refer to a unique number of visitors that have reached the website for a given period of time. CR also called conversion rate is the percentage of visitors to the website that completes the desired goal (conversion) out of the total number of visitors. A higher CR indicates that a greater number of users get converted that landed onto the website. Path length is the average length of the channel path.

5.4 Technology and Frameworks

Python and *R* were used as the main programming language for running a different kind of experiment. ETL operations were written in Python and some modelling process were conducted in R along with data visualizations. There are a bunch of libraries for both Python and R, but the major ones are included in this section.

Python

- *Numpy (version == 1.14.2)* is used for matrices operations
- *Pandas (version == 0.22.0)* is used for data frame operations and data aggregation
- *Scikit-learn (version == 0.19.1)* is used for different machine learning algorithms such as logistic regression
- *Imbalanced-learn (version == 0.3.3)* is used for handling imbalanced class data.
- *Matplotlib (version == 2.2.2)* is used for drawing plots

R

- *Markov chain (version == 0.6.9.10)* is used to apply markov chain on sequence data.
- *Dplyr (version == version 0.7.5)* is used to perform data aggregation by writing SQL like queries.
- *Shiny (version == 0.10.2.2)*. It is a package to build interactive web applications with data visualization in the form of a dashboard.

5.5 Evaluation Metrics

Variability Measure

Standard deviation is computed for individual coefficients and then it is averaged across all campaigns. Standard deviation shows how far it is from the mean. It gives an estimate of stable coefficients and the variability in them. A lower standard deviation gives an idea that the coefficients across different campaigns over 10-fold iterations do not change and they are stable coefficients. [Shao and Li, 2011]

Confusion Matrix

Table 5.2 shows the confusion matrix that tells the performance of a classification model for a set of values of test set for which true values are known. [Sun *et al.* 2009]

N = sample size	**Predicted: NO**	**Predicted: YES**
Actual: NO	**TN**	**FP**
Actual: YES	**FN**	**TP**

Table 5.2 Confusion Matrix

True Negatives (TN): The truth value was NO (not converted) and the model also predicted it as NO (not converted).

True Positives (TP): The truth value was YES (converted) and the model also predicted as YES (converted).

False Positives (FP): The truth value was NO (not converted) but the model predicted as YES (converted). It is also known as *the Type I error*.

False Negatives (FN): The truth value was YES (converted) but the model predicted as NO (not converted). It is also known as *the Type II error*.

Misclassification rate

Misclassification rate is also known as *the error rate*. It tells overall, how often the prediction is wrong. It can be calculated as:

$$Misclassification\ rate = \frac{FP+FN}{total},$$

where the total is the sample size.

Recall

Recall is also known as True Positive Rate. It tells when it is actually YES (converted), how often does it predict YES (converted). It can be calculated as:

$$Recall = \frac{TP}{TP+FN},$$

[Sun *et al.* 2009]

ROC AUC

ROC AUC is a commonly used metric to visualize the performance of a binary classifier. It summarizes the performance of a classifier over all possible thresholds. ROC curve Sun *et al.* [2009] plots *True Positive Rate* (y-axis) against *False Positive Rate* (x-axis) over varying thresholds. AUC score is then calculated by taking the area under the curve.

Precision

Precision shows when it predicts YES (converted), how often is it correct.

$$Precision = \frac{TP}{Predicted\ YES},$$

[Sun *et al.* 2009]

F-measure

F-measure Sun *et al.* [2009] is a weighted average of precision and recall. It conveys the balance between precision and recall.

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

6. PROPOSED METHODS

In order to address the problem of attribution modelling, the methods used should yield such results that support decision making. Marketers face an optimization problem as to how to optimize their campaigns in order to achieve highest ROI. Attribution modelling should be able to address this same problem while maintaining the interpretability of the models which marketers can easily use. This helps in allocating budget to each campaign and optimizing their marketing strategies. In principle, it means to measure the effect of individual marketing channels contributing to the conversion process. For the sake of analysis, two approaches have been taken into consideration to estimate the contribution of each channel leading up to conversion i.e. user level data and sequence level data.

6.1 Logistic Regression

Logistic regression was used on user level data. The process should be easy to interpret, and the outcome should be able to address the importance of each marketing channel in relation to others [Shao and Li, 2011]. Logistic regression was used to estimate the probability that the user would convert or not based on a set of features within the time period of the analysis. It was a binary classification task with conversion value being 0 or 1. 0 means not converted while 1 means converted. The independent variables are marketing channels while the dependent variable is conversion. The process can be defined according to the following equation.

$$p(conversion) = \beta_0 + \beta_1 C1 + \beta_2 C2 + \cdots,$$

where $C1$ and $C2$ are marketing channels and β_o is the y-intercept.
With y-intercept being removed, the coefficients are then extracted for each channel. These coefficients are then used for attribution modelling.

6.1.1 Attribution Modelling

Digital attribution is the allocation of conversions onto the marketing channels through which online advertisements are served to the audience [Katsov, 2017]. A correct attribution is a necessity to be able to evaluate the efficacy of the channels, which is, in turn, the basis for re-allocation of funding into the channels to improve the efficiency of advertisement. The process is outlined as follow.

- Weights are allocated to each user for each channel that leads up to conversion. Weight allocation is done by multiplying each user interaction with a channel with

the coefficient of that channel. It can be expressed according to the following equation.

$$converters = U_j C_i * \beta_i,$$

Where $U_j C_i$ denotes the count of an interaction of user *j* with advertising channel *i* and β_i, *i =1….. p* denotes *p* different advertising channels.

- With these weights allocated, they are then transformed into *credits* which describe an amount of conversion credit.
- Assume there are 4 channels as in the table, *a, b, c, d.*
- The credit assigned to these channels is the decrease in the probability of conversion that occurs for each converting user *i* after removing the channel from the converting path.
- So, if $\boldsymbol{x_i} = (\boldsymbol{n_{i_a}}, \boldsymbol{n_{i_b}}, \boldsymbol{n_{i_c}}, \boldsymbol{n_{i_d}})$, we calculate the drop in event probability for $\boldsymbol{x_{i_{-a}}} = (\boldsymbol{0}, \boldsymbol{n_{i_b}}, \boldsymbol{n_{i_c}}, \boldsymbol{n_{i_d}})$, and so for all the other subchannels.

$$Credit(a) = \sum_i f(x_i) - f(x_{i-a}) = \sum_i \frac{1}{1 + e^{-(\alpha + \beta^t x_i)}} - \frac{1}{1 + e^{-(\alpha + \beta^t x_{i-a})}}$$

- We only account for the credit in paths that lead up to conversion.
- The credits are then normalized for each converter to sum to 1.
- The total credit for all channels, therefore, equals the number of converting users.
- The credits are then aggregated for each channel. Since there are two separate channels for clicks and impression, final conversion credits are obtained by aggregating credits for clicks and impressions.

Channels that consistently appear in pathways leading to a conversion get assigned a higher weight than channels that do not appear as often in those pathways. The basis of the attribution is the difference of the probabilities for the reduced pathways to the probability of conversion for the full path. The decrease in probability equals the amount of conversion credit a channel received in the conversion path. They reflect how much each channel helped to convert that particular user. Finally, the conversion is attributed to each channel proportional to the difference in probabilities. This process is repeated for every conversion, which gives us, in the end, the sum of conversions that each channel generated. Using this weighting all conversions get attributed to the channels in their pathways, with each touch point getting a fraction of the credit proportional to the weight assigned by the statistical model. This method of attribution provides a realistic picture of the efficacy of each channel, which in turn allows to compute and compare each channel's efficiency used to deliver advertisements.

Since the dataset is highly imbalanced with not converted being in majority, different sampling methods were used to address the class imbalance problem in order to improve the predictive power of logistic regression and stability of coefficients.

Random undersampling

In this sampling technique, samples from the majority class (not converted) were picked at random from the data and were removed. After this, the distribution of samples from both classes was evenly balanced at a ratio of 50-50. [Sun *et al.* 2009]

SMOTE

Synthetic minority over sampling technique (SMOTE) addresses the class imbalance problem by creating new synthetic samples of the minority class (conversion). The minority class is up sampled to balance the majority class on a ratio of 50-50. SMOTE creates new minority instances between real minority instances as shown in Figure 6.1. "The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. The implementation currently uses five nearest neighbors. For instance, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1 and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general". [Chawla *et al.* 2002]

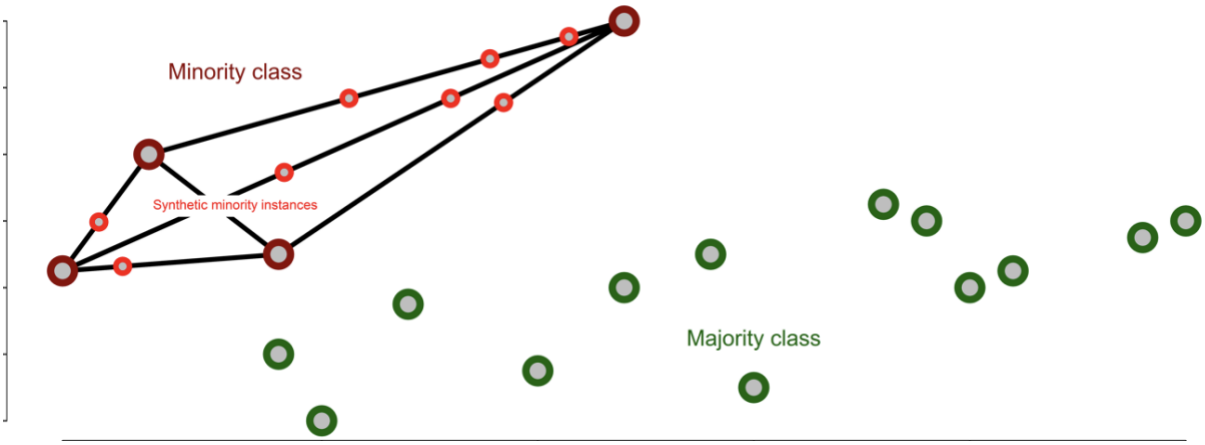

Figure 6.1 SMOTE [Kunert, 2017]

Bagging

Bagging is an ensemble technique used to reduce variance in the data by combining multiple classifiers which are trained on different subsamples of same data. The predictions from multiple classifiers are then combined together to give a better classifier with less variance. A previous research [Shao and Li, 2011] showed that bagging yielded more stable coefficients so multiple logistic regression models were trained on subsamples of user-level data and the coefficients were averaged out to provide more stable coefficients. [Breiman, 1996]

In supervised machine learning such as classification, models are trained on a training data. This often leads to overfitting. Overfitting means that the model doesn't generalize well for the training data to unseen data. This leads to poor model performance on unseen data. Regularization addresses the overfitting problem by making the models simpler. It adds a penalty term to the fitting error. Two regularization techniques *L1* and *L2* were used with logistic regression to avoid overfitting.

L1 Regularization

L1 regularization Lee *et al*. [2006] adds absolute value of magnitude of coefficient as penalty term to the loss function. L1 regularized logistic regression can be defined as:

$$L1\ Regularized\ Logistic\ Regression = \sum_{n=0}^{N-1} \ln\left(1 + exp^{y_n w^T x_n}\right) + \lambda\ ||\mathrm{w}||_1,$$

where $||\mathrm{w}||_1$ is L1 penalty, $\ln\left(1 + exp^{y_n w^T x_n}\right)$ is the log-loss and *w* are the weight vectors.

L2 Regularization

L2 regularization [Moore and DeNero 2011] adds the squared magnitude of coefficient as the penalty to the loss function. It can be defined as:

$$L2\ Regularized\ Logistic\ Regression = \sum_{n=0}^{N-1} \ln\left(1 + exp^{y_n w^T x_n}\right) + \lambda\, w^T\, w,$$

where $w^T w$ is L2 penalty, $\ln\left(1 + exp^{y_n w^T x_n}\right)$ is the log-loss and *w* are the weight vectors.

Log loss also known as logarithmic loss measures the performance of a classification model based on prediction probabilities as input. The goal is to minimize this value. Log loss increases as L1 produces sparse outputs so inherently supports feature selection producing many coefficients with zero values and few large coefficients. L1 is the sum of weights while L2 is the sum of the square of weights.

Stratified Cross-validation

Cross-validation Krstajic *et al.* [2014] is a resampling technique which is used to prevent the model from overfitting. It shuffles the data set randomly and split the dataset into K folds. For each k^{th} fold, the k^{th} fold is treated as a test set and $k^{th} - 1$ folds as the training set. In this way, the model is trained and evaluated k^{th} times and prediction are averaged out. It would a less biased estimate. Since the dataset is highly imbalanced, the not converted class being in majority so K-fold cross-validation was done with stratification commonly known as stratified cross-validation. In this way, in each fold, the distribution of classes remains the same and data is not shuffled randomly.

6.2 Markov Chain

Markov chain was used with sequence level data where each row of the data represents a consumer journey leading up to a conversion or not conversion. The problem can be interpreted as a markov process [Katsov, 2017] in which the future is independent of the past, given the present. In the case of attribution modelling, each sequence represents a touch point along the consumer journey in increasing order of time. Each sequence was represented as a directed graph with each vertex as a touch point called *state* and edges represent the probability of moving from one state to another. p_{ij} is the probability of

moving from $state_i$ to $state_j$. The probabilities p_{ij} are called transition probabilities. The starting state S is the first touch point in the customer journey with two ending states as conversion (1) or not conversion (0). The markov process can be defined as:

$$\mathbb{P}(X_{t+1} = s \mid X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = \mathbb{P}(X_{t+1} = s \mid X_t = s_t),$$

where X_t is the state of Markov chain at time *t*, for all *t* = 1,2, 3, …. and for all states $s_o, s_1, \ldots s_t, s$.

The transition probability p_{ij} can be defined as:

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i).$$

By computing the transition probabilities, conversion credits can be attributed to each channel showing the impact of the channel in user conversion. Markov graphs are easy to interpret and show a comprehensive view of consumer journeys. In this thesis, *the first-order markov chain* was used to attribute conversion credits to channel. In first-order markov chain each state is dependent only on the previous one. It can be explained according to the following conditional probability.

$$p(x_t|x_{t-1}),$$

where x_t is the current state.

6.2.1 Attribution Modelling

In the context of markov chain, the conversion credits are allocated to each channel using the *removal effect* Anderl *et al.* [2013]. Each channel is consecutively removed from the graph and measure how many conversions could be made in the absence of the channel. Removal effect of channel *k* can be defined as follows:

$$Removal\ effect\ of\ channel\ k = \frac{p(conversion\ in\ absence\ of\ channel\ k)}{p(conversion\ in\ presence\ of\ channel\ k)},$$

where k = 1…N, N is the number of channels.

$$p(conversion) = \sum_{n=1}^{N} \prod p_{ij},$$

where N is the number of converting paths and p_{ij} is the probability of moving from state i to state j.

Conversion credits can then be calculated as:

$$Credits(k) = Removal\ effect\ of\ channel\ k * total\ conversions,$$

where k= 1…N, N is the number of channels.

This can be further explained by looking at Figure 6.2.

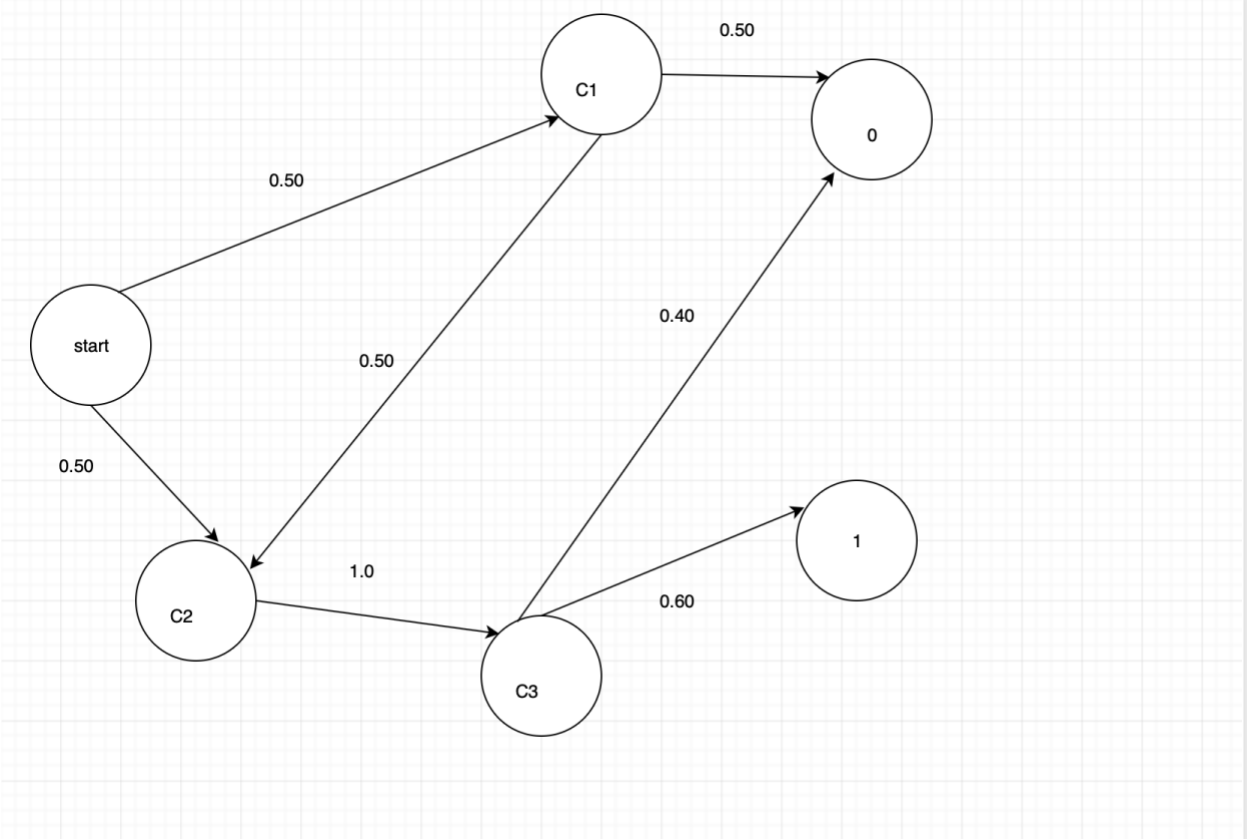

Figure 6.2 Markov chain graph with channels

Figure 6.2 shows three channels C1, C2, C3 with their state transition probabilities represented by edges. 0 means, not conversion and 1 means conversion. Now removal effect would be used to estimate the conversion credits for C1 which in turn would show the contribution of C1 in conversion.

$$p(conversion) = P(C1 \rightarrow C2 \rightarrow C3 - 1) + p(C2 \rightarrow C3 \rightarrow 1)$$

$$p(conversion) = 0.5 * 0.5 * 1.0 * 0.6 + \ 0.5 * 1 * 0.6$$

$$p(conversion) = 0.45$$

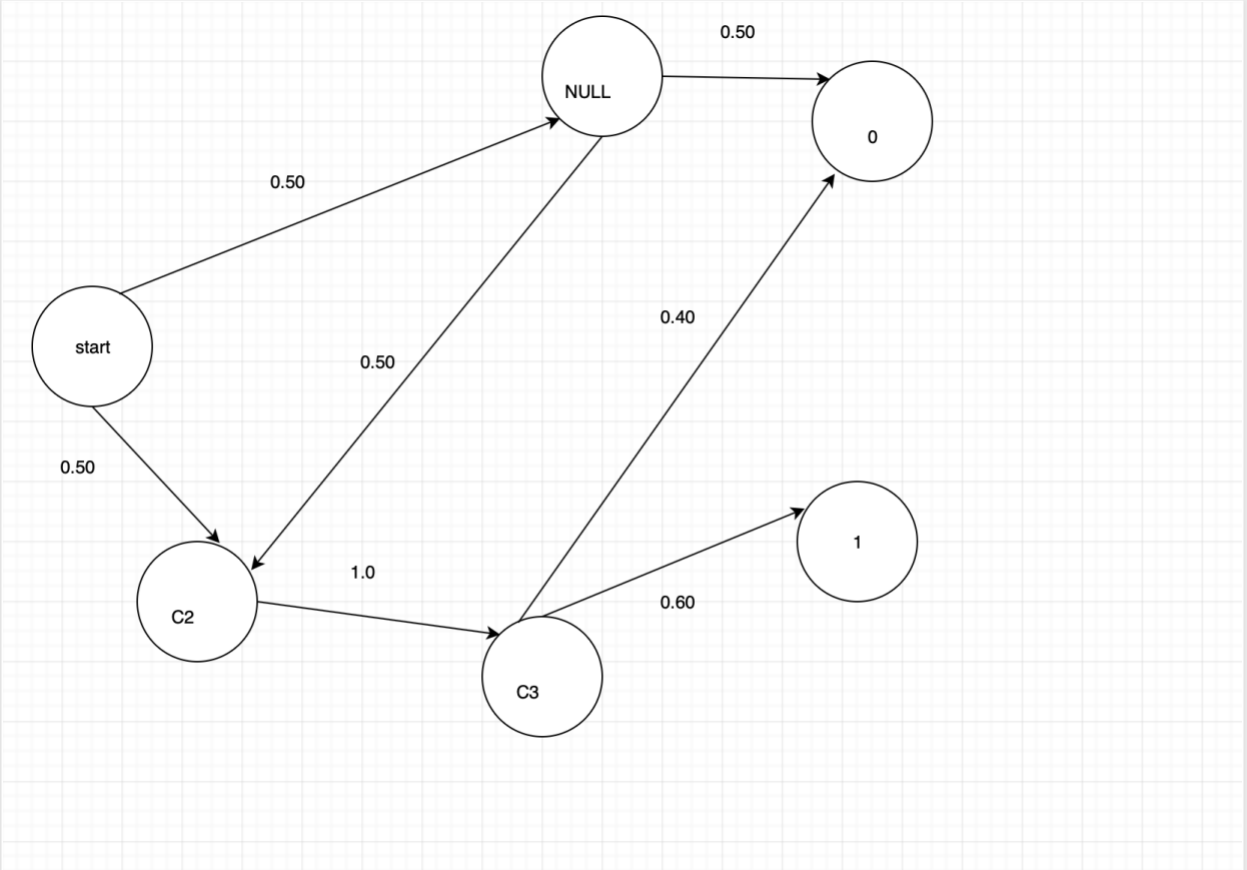

Figure 6.3 Markov chain graph with C1 channel removed

In Figure 6.3, channel C1 has been removed and replaced with the NULL state. Now conversion probability is calculated in absence of C1.

$$p(conversion\ without\ C1) = p(C2 \rightarrow C3 \rightarrow 1)$$

$$p(conversion\ without\ C1) = 0.5 * 1.0 * 0.6$$

$$p(conversion\ without\ C1) = 0.30$$

Removal effect of C1 can be calculated as:

$$Removal\ effect\ of\ C1 = \frac{0.30}{0.45}$$
$$= 0.67$$

It means that channel C1 contributes to 67% of conversions in consumer journey. The process is repeated for every channel in the Markov graph and contribution of each channel is then estimated.

7. EXPERIMENTS AND RESULTS

7.1 Logistic Regression

In order to address the problem, logistic regression was used with variations on user level data. Logistic regression is easy to interpret which is important for the attribution modelling to be used in practice. All evaluation metrics are computed over 10-fold stratified cross-validation.

7.1.1 Random Undersampling

Sun *et al.* [2009] address multiple techniques to address the problem of an imbalanced dataset. One such solution is random undersampling. Due to the imbalanced nature of class distribution with the non-conversion class being the majority class, random undersampling was used to balance the class distribution. In the random undersampling, the data points from the majority class (non-conversion) are removed at random until the distribution of the 2 classes is evenly balanced at a ratio of 50-50. Logistic regression was then used on this balanced data set. Table 7.1 summarizes the results in the form of a confusion matrix. The confusion matrix was computed on 80% training data and 20% test data. All results are aggregated over 10-fold stratified cross validation.

	Predicted = Non-Conversion	Predicted = Conversion	
Actual= Non-Conversion	45 (0.43)	7 (0.06)	52
Actual= Conversion	35 (0.32)	21 (0.19)	56
	80	28	

Table 7.1 Confusion Matrix after random undersampling

In Table 7.1, 43% of the samples were True Negative. 6% of the samples were False Positive. 32% of the samples were False Negative and 19% of the samples were True Positive.

Variability Measure	0.20
Misclassification rate	0.39
Recall	0.38
ROC AUC	0.66

Table 7.2 Metrics computed after random undersampling

In Table 7.2, the metrics were computed to assess the performance of logistic regression after doing random undersampling.

7.1.2 Synthetic Minority Oversampling

Another sampling technique was used to balance the class distribution. Synthetic minority oversampling technique (SMOTE) Chawla *et al.* [2002] balance the class distribution by increasing positive (conversion) samples. The classes were evenly balanced at a ratio of 50-50. The confusion matrix was computed on 80% training data and 20% test data. All results are aggregated over 10-fold stratified cross validation.

	Predicted = Non-Conversion	Predicted = Conversion	
Actual= Non-Conversion	571 (0.44)	70 (0.05)	641
Actual= Conversion	316 (0.25)	331 (0.26)	647
	887	401	

Table 7.3 Confusion Matrix after SMOTE

In Table 7.3, 44% of the samples were True Negative. 5% of the samples were False Positive. 25% of the samples were False Negative and 26% of the samples were True Positive.

Variability Measure	0.51
Misclassification rate	0.29
Recall	0.51
ROC AUC	0.74

Table 7.4 Metrics computed after SMOTE

Table 7.4 shows the metrics that were computed to assess the performance of logistic regression after doing upsampling using SMOTE.

7.1.3 Weighted Logistic Regression

Since the data was highly imbalanced, weights were added to classes Sun *et al.* [2009]. Since our class of interest is the minority class (conversion) so a higher weight was assigned to minority class as opposed to the majority class. The distribution of class was computed from the dataset and each class was assigned the weight as the proportion of other class in the dataset as shown in Table 7.5.

Class	**Class weight**
Conversion	0.92
Non-Conversion	0.08

Table 7.5 weights assigned to classes

The experiment was done on 80% training data and 20% test data. Since the distribution of class is unchanged, miss-classification rate was not computed for this experiment as the data is not balanced. All results are aggregated over 10-fold stratified cross validation. The confusion matrix is computed as below:

	Predicted = Non-Conversion	Predicted = Conversion	
Actual= Non-Conversion	284 (0.82)	32 (0.09)	316
Actual= Conversion	17 (0.05)	15 (0.04)	32
	301	47	

Table 7.6 Confusion Matrix of Weighted Logistic Regression

In Table 7.6, 82% of the samples were True Negative. 9% of the samples were False Positive. 5% of the samples were False Negative and 4% of the samples were True Positive.

Variability Measure	0.13
Recall	0.43
ROC AUC	0.65

Table 7.7 Metrics of Weighted Logistic Regression

Table 7.7 shows the metrics that were computed to assess the performance of logistic regression after adding weights to class samples.

The experiment was repeated by adding bagging [Shao and Li, 2011] which is a special case of model averaging. It helps reduce variance in the data and avoids overfitting.

Variability Measure	0.13
Recall	0.43
ROC AUC	0.66

Table 7.8 Metrics of Weighted Logistic Regression after Bagging

Table 7.8 shows the metrics that were computed to assess the performance of *weighted bagged logistic regression.*

The experiment was repeated with different regularization *L*1 Lee *et al*. [2006] and *L2* [Moore and DeNero, 2011] for both weighted logistic regression and bagged weighted logistic regression. Both the methods yielded similar results.

Variability Measure	0.05
Recall	0.42
ROC AUC	0.65

Table 7.9 Metrics after adding L1 regularization

Table 7.9 shows the metrics that were computed after adding L1 regularization.

Variability Measure	0.08
Recall	0.43
ROC AUC	0.69

Table 7.10 Metrics after adding L2 regularization

Table 7.10 shows the metrics that were computed after adding L2 regularization.

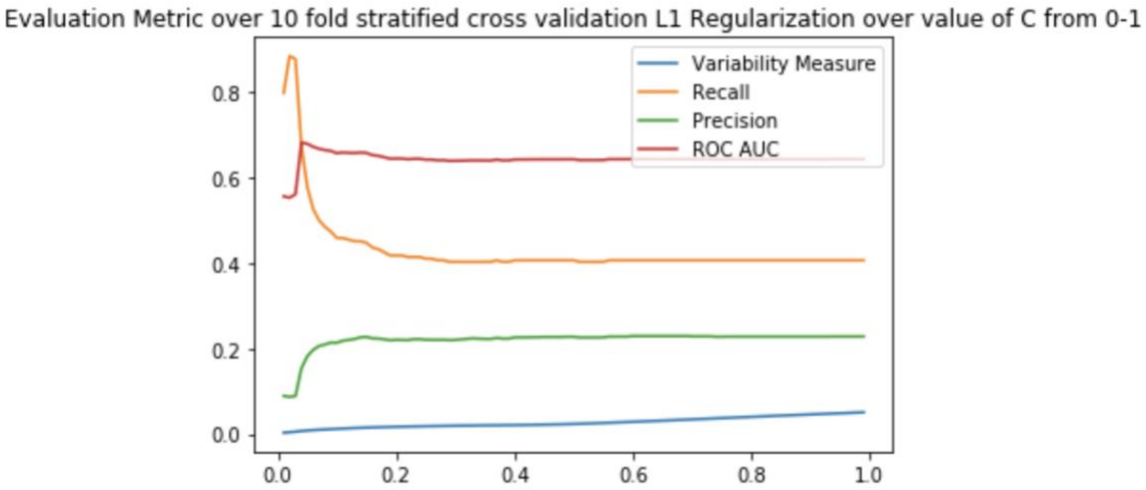

Figure 7.1 L1 Metrics over 10-fold stratified cross-validation

In Figure 7.1, metrics are computed over 10-fold stratified cross-validation using L1 regularization over a range of C values. C is inverse of regularization. Smaller values of C indicate stronger regularization. The experiment is conducted on weighted logistic regression.

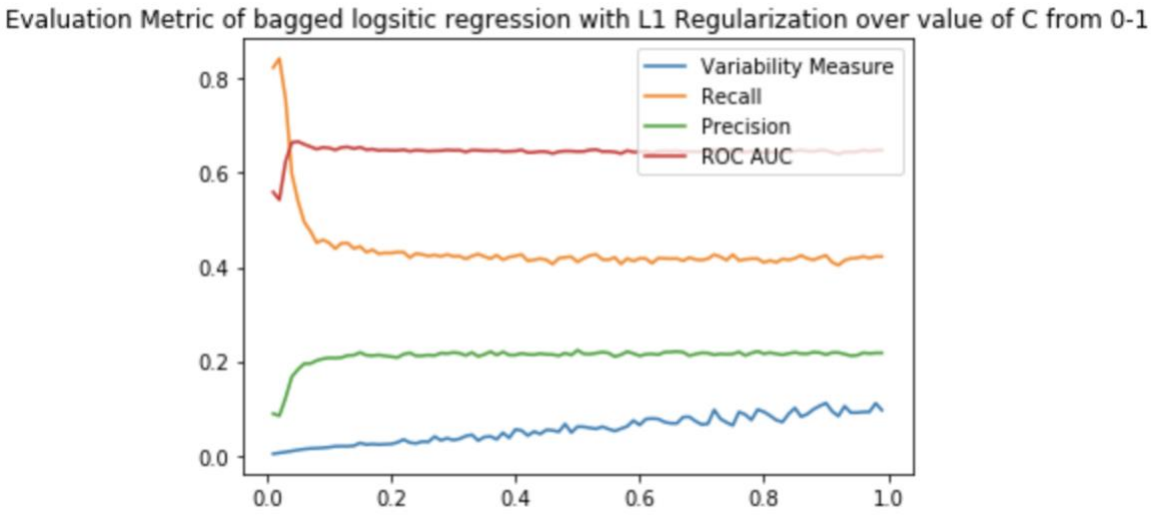

Figure 7.2 L1 Metrics over 10-fold stratified cross-validation using bagging

In Figure 7.2, the experiment is repeated using weighted bagged logistic regression.

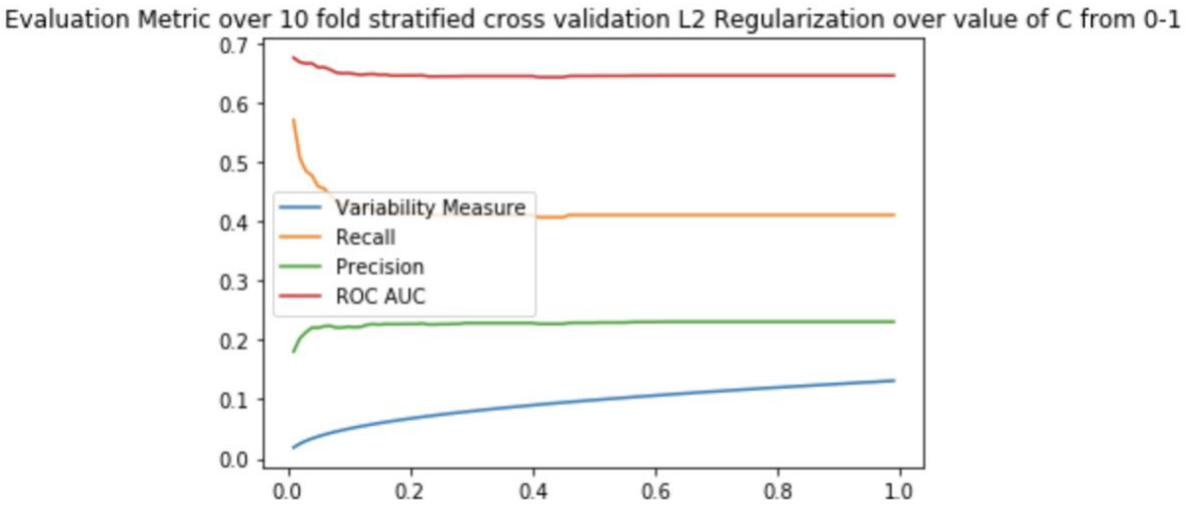

Figure 7.3 L2 metrics over 10-fold stratified cross-validation

In Figure 7.3, metrics are computed over 10-fold stratified cross-validation using L2 regularization over a range of C values. C is inverse of regularization. Smaller values of C indicate stronger regularization. The experiment is conducted on weighted logistic regression.

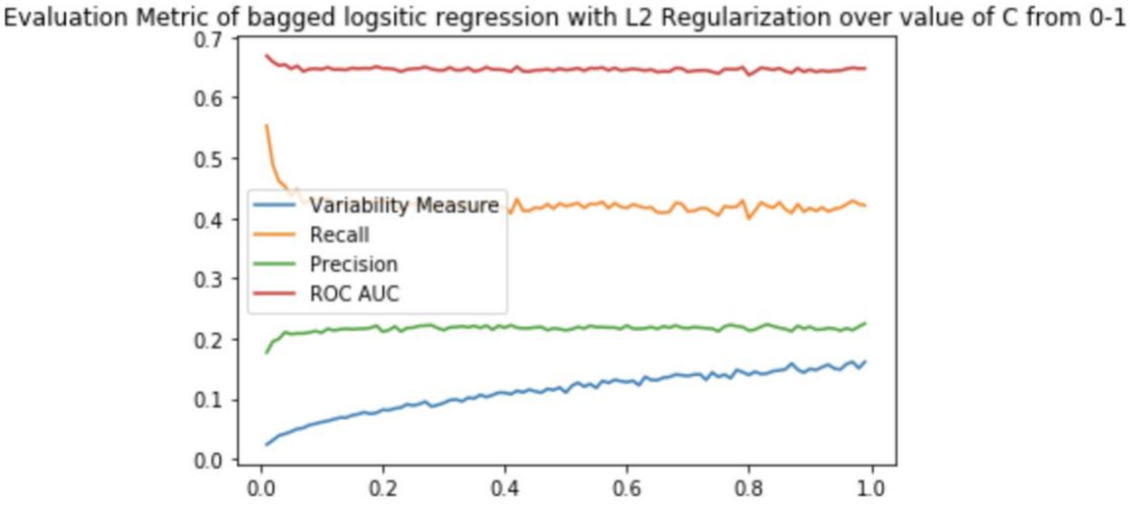

Figure 7.4 L2 metrics over 10-fold stratified cross-validation using bagging

In Figure 7.4, the experiment is repeated using weighted bagged logistic regression.

Methods	Variability Measure	ROC AUC	Recall
LR with Random Undersampling	0.20	0.66	0.38
LR with SMOTE	0.51	0.74	0.51
LR with weights	0.13	0.65	0.43
LR with weights and Bagging	0.13	0.66	0.43
LR with weights, and L1 regularization	0.05	0.65	0.43
LR with weights, and L2 regularization	0.05	0.68	0.42
LR with weights, Bagging and L1 regularization	0.03	0.65	0.43
LR with weights, Bagging and L2 regularization	0.07	0.68	0.42

Table 7.11 Comparable results of methods used in user-level data.

Table 7.11 shows comparable results of all the methods used in user level data. It indicates that weighted logistic regression with and without bagging yielded similar results. However, it outperformed logistic regression with random undersampling and SMOTE. In terms of ROC AUC, weighted logistic regression with L2 regularization outperformed L1 regularization but suffered from slightly low variability measure. Overall due to imbalance nature of data, random undersampling and SMOTE didn't prove to be useful in terms of variability measure. On the other hand, adding class weights to logistic regression with imbalanced data gave improved results in terms of variability measure and ROC AUC.

7.2 Markov Chain

Markov chain makes use of sequence level data to attribute credits to conversions. Due to its probabilistic nature, markov chain is easier to interpret as one can easily visualize the states and how the user would shift from one state to another using a state transition matrix. All results are aggregated over 10-fold stratified cross validation. The results using first-order markov chain are summarized below:

	Predicted = Non-Conversion	Predicted = Conversion	
Actual= Non-Conversion	3244 (0.64)	143 (0.03)	3387
Actual= Conversion	484 (0.09)	1247 (0.24)	1731
	3728	1390	

Table 7.12 Confusion Matrix of First-order Markov Chain

In Table 7.12, 64% of the samples were True Negative. 3% of the samples were False Positive. 9% of the samples were False Negative and 24% of the samples were True Positive.

Recall	0.90
Precision	0.72
F-measure	0.80

Table 7.13 Metrics computed using Markov Chain

Table 7.13 shows the metrics that were computed to assess the performance of first-order markov chain.

	0	1	D1	D2	S1	S2	D3	D4	D5	D6	D7	D8
0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0
D1	0.008	0	0.99	0	0	0	0	0	0	0	0	0
D2	0.018	0.038	0	0.90	0	0.0018	0.007	0	0.0079	0.006	0.02	0
S1	0.33	0	0	0	0.67	0	0	0	0	0	0	0
S2	0.10	0.006	0	0.00015	0	0.88	0.002	0.0003	0.0004	0.006	0.004	0.0002
D3	0.02	0.002	0	0.0007	0	0.00002	0.96	0.0002	0.005	0.003	0.001	0.0002
D4	0.167	0	0	0	0	0.0006	0.004	0.98	0.002	0	0	0
D5	0.066	0	0.00006	0.002	0	0.0001	0.014	0.0002	0.91	0.007	0	0
D6	0.02	0.001	0	0.002	0	0.002	0.01	0.00006	0.01	0.95	0.002	0.0002
D7	0.01	0.04	0	0.005	0	0.001	0.006	0	0.003	0.003	0.93	0.0004
D8	0.02	0.007	0	0.004	0	0.004	0.02	0	0.004	0.0087	0.01	0.92

Table 7.14 Markov Chain transition matrix

Table 7.14 shows a state transition matrix for the first order Markov chain. The values are state transition probabilities, probability of moving from one state to another. The states are campaigns with D1 as Display 1 to D8 as Display 8. The social media campaigns are S1 as Social Media 1 and S2 as Social Media 2.

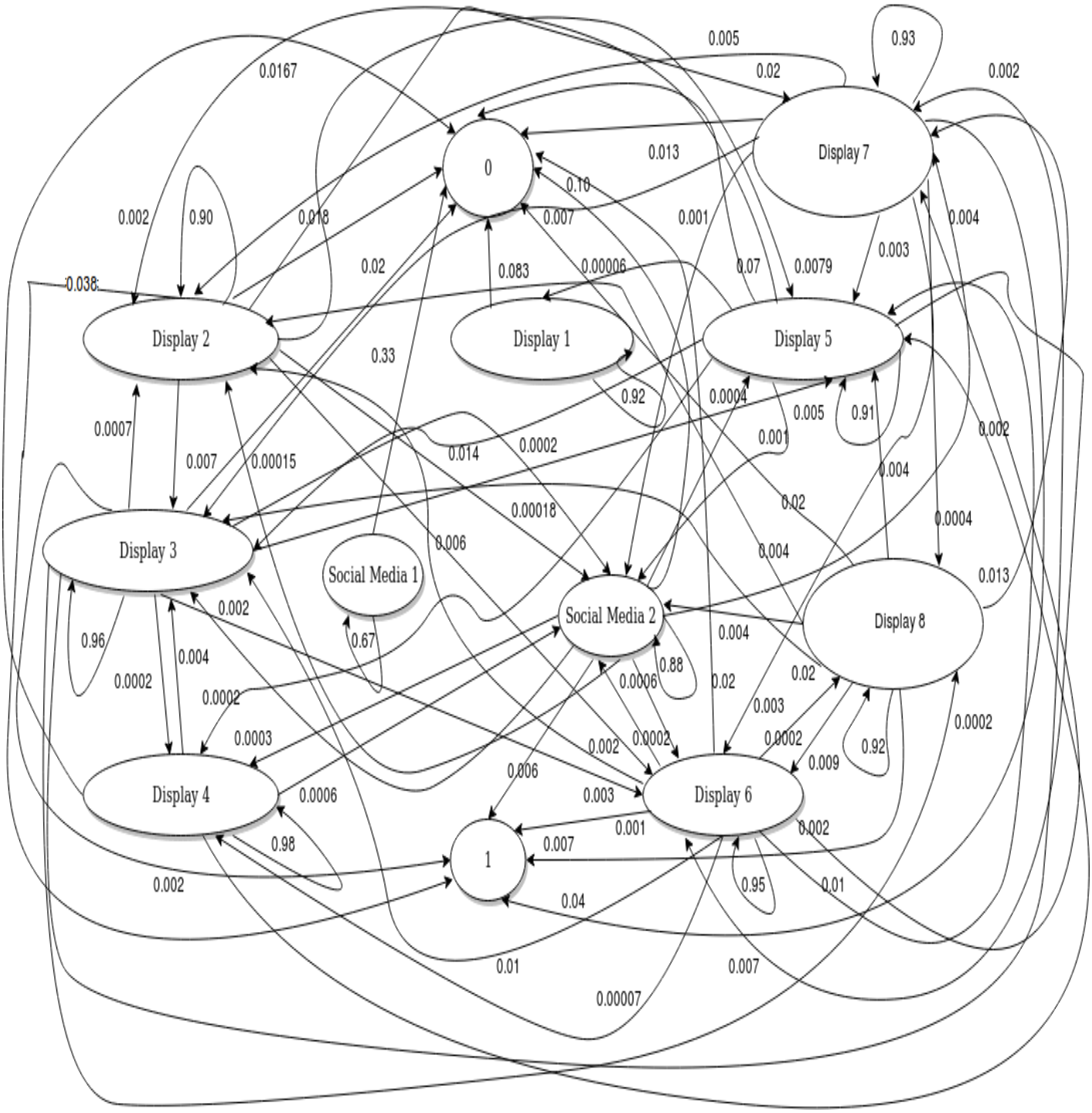

Figure 7.5 Markov chain graph with all the channels

Figure 7.5 shows the markov chain graph with all the customer interactions and state transition probabilities leading up to conversion as state 1 and non-conversion as state 0.

8. CONCLUSION

The project was done in collaboration with Annalect Finland. The campaign names were anonymized as to hide the identity of the customer. Only one week of campaigns data was used which can be easily processed on local machines. The thesis required sufficient domain knowledge of digital marketing which was possible though with the help of co-workers.

The dataset had its fair share of challenges. Since there were two data sources for attribution modelling, Adform and Google Analytics, much of the time went into data aggregation and preparing for the models. The challenge was to understand how digital marketing works and how the campaigns run in a digital medium. Moreover, familiarity with terms like CTR, impressions, clicks etc. was necessary to find the underlying correlation between features. The imbalance nature of data added increased complexity to model the problem in such a way that it can generalize well.

Two different approaches were used to address attribution modelling. In the first problem, the data was prepared in such a way which can be treated as a classification task. Logistic regression was used with different regularization techniques and bagging to achieve state of the art results and reduce overfitting due to high number of covariates [Shao and Li, 2011]. As pointed out in Table 7.11, bagging, weights and L1 regularization yielded a significant impact on variability measure and stability of coefficients. Bagging yielded variability measure of 0.03 as compared to [Shao and Li, 2011] results pointed out in Table 4.1 which yielded a variability measure of 0.672. Weighted logistic regression performed best with regard to predictive accuracy and addressing the class imbalance problem. The tradeoff between stability of coefficients and predictive accuracy of the model was best addressed by a combination of logistic regression with weights, bagging and L1 regularization yielding a variability measure of 0.03, ROC AUC of 0.65 and recall of 0.43.

The second problem models the data as a markov process. This increased the data aggregation task as data format widely differs for both the problems. However, it gave an alternating viewpoint addressing the same problem. Markov process adds increased interpretability to the attribution framework Anderl *et al.* [2013] as compared to logistic regression due to its state transition matrix which maps customer journeys in the form of graphs and probability of moving from one state to another. It helps the marketers to readily adopt it as to how to devise their marketing strategies. First order markov chain was used which outperformed logistic regression in respect of predictive accuracy. The model was able to obtain a recall of 0.90, precision of 0.72 and F-measure of 0.80 as pointed out in Table 7.13. In Figure 7.5, Markov chain graph gives a comprehensive view of customer journeys which helps the marketers to plan budget allocation as to which channel is driving most of the user conversions.

In a nutshell, as opposed to popular "predictive modelling", it was interesting to work with attribution modelling which is a use case very specific to digital marketing.

References

ACMer. 2016. A Short Introduction - Logistic Regression Algorithm. Unpublished manuscript, March 17, 2016. Available as https://helloacm.com/a-short-introduction-logistic-regression-algorithm/. Checked August 21, 2018.

V. Abhishek, P. Fader, K. Hosanagar. Media Exposure through the Funnel: A Model of Multi-Stage Attribution. SSRN, October 2012.

E. Andrel, I. Becker, F. V. Wangenheim, and J. H. Schumann, 2013. Putting attribution to work: A graph-based framework for attribution modeling in managerial practice. Unpublished manuscript, October 23, 2013. Available as https://ssrn.com/abstract=2343077. Checked December 26, 2018.

L. Breiman. 1996. Bagging predictors. *Machine Learning* 24(2), 123-140.

N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321-357

J. Con. 2016. All 11 Marketing Attribution Models, Explained. Unpublished manuscript, July 15, 2016. Available as https://www.bizible.com/blog/marketing-attribution-models-complete-list. Checked November 10, 2018.

D. R. Cox. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society* 20(2), 215-242.

D. E. Daykin, J. E. Jeacocke and D. G. Neal. 1967. Markov chains and snakes and ladders. *The Mathematical Gazette* 51(378), 313-317.

R. W. Gehl. 2016. *Socialbots and their Friends: Digital Media and the Automation of Sociality.* M. Bakardjieva, 2016.

A. Goldfarb, C. Tucker. 2011. Online display advertising: targeting and obtrusiveness.

Marketing Science 30(3), 389-404

C. M. Grinstead. 1988. *Introduction to Probability.* J. L. Snell, 1988.

B. J. Jansen, S. Schuster. 2011. Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research* 12(1), 1-18.

C. H. W. Jayawardane, S. K. Halgamuge, U. Kayande. 2015. Attributing Conversion Credit in an Online Environment: An Analysis and Classification. *In: Proc. of the 3rd International Symposium on Computational and Business Intelligence (ISCBI). IEEE,* 68-73.

Dr. R. Kunert. 2017. SMOTE explained for noobs - Synthetic Minority Over-sampling Technique line by line. Unpublished manuscript, November 06, 2017. Available as http://rikunert.com/SMOTE_explained. Checked November 3, 2018.

I. Katsov. 2017. *Introduction to Algorithmic Marketing.*

D. Krstajic, L. J. Buturovic, D. E. Leahy and S. Thomas. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* 6(10), 15 pages.

S. Lee, H. Lee, P. Abbeel and A. Y. Ng. 2006. Efficient L_1 Regularized Logistic Regression. *In: Proc. of the Twenty-First National Conference on Artificial Intelligence (AAAI-06) and Eighteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-06),* 401-408.

R. Manjur. 2018. A breakdown of global ad spends across mediums. Unpublished manuscript, January 12, 2018. Available as http://www.marketing-interactive.com/a-break-down-of-global-ad-spend-across-mediums/. Checked August 21, 2018.

R. Moore, J. DeNero. 2011. L1 and L2 Regularization for Multiclass Hinge Loss

Models. *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, 1-5.

N. Necsulescu. 2015. Multi-channel attribution and its role in marketing investment. *Journal of Digital & Social Media Marketing* 3(2), 125-134.

K. Neville. 2017. *Channel attribution modelling using clickstream data from an online store*, Master's thesis, Statistics and Data Mining, Linkoping University.

L. Page, S. Brin, R. Motwani and T. Winograd. 1999. The PageRank Citation Ranking: Bringing order to the web. Unpublished manuscript, January 29, 1998. Available as http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf. Checked February 5, 2019.

C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll Indiana University- Bloomington. 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96(1), 3-14.

V. Powell, L. Lehe. 2014. Markov Chains Explained Visually. Unpublished manuscript, November 07, 2014. Available as http://setosa.io/ev/markov-chains/. Checked November 5, 2018.

PwC. 2018. IAB internet advertising revenue report 2017 full year results in an industry survey conducted by PwC and sponsored by the Interactive Advertising Bureau (IAB). Unpublished manuscript, May 2018. Available as https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-Report.REV_.pdf. Checked August 21, 2018. Checked August 21, 2018.

O. Rentola. 2014. *Analyses of Online Advertising Performance Using Attribution Modeling*, Master's thesis, School of Science, Aalto University

S. J. Russel, P. Norvig. 2009. *Artificial Intelligence: A Modern Approach (AIMA).*

A. L. Samuel. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3(3)*,* 210-229.

S. Sapp, J. Vaver, M. Shi, and N. Bathia. DASS: Digital Advertising System Simulation. http: //research.google.com/pubs/ pub45331.html, 2016.

X. Shao and L. Li. Data-driven multi-touch attribution models. 2011. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 258–264.

Y. Sun, A. K. C. Wong and M. S. Kamel. 2009. Classification of imbalanced data: a review*. International Journal of Pattern Recognition and Artificial Intelligence* 23(4), 687-719.

L. L. Weiss, 1964. Sequences of wet or dry days described by a markov chain probability model. *Monthly Weather Review* 92(4), 169-176.

M. M. Yadagiri, S. K. Saini, R. Sinha. 2015. A Non-parametric Approach to the Multi-channel Attribution Problem. In: *Proc. of the 16th International Conference on Web Information Systems Engineering, Springer,* 338-352*.*

K. Zhao, S. H. Mahboobi, S. R. Bagheri, 2018. Revenue-based attribution modeling for online advertising. *International Journal of Market Research*, 21 pages.